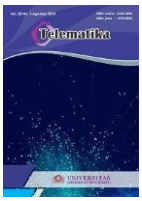

Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemenristek/BRIN, No. 85/M/KPT/2020

UNIVERSITAS AMIKOM PURWOKERTO

Performance Analysis of Ensemble Learning Models in Heart Failure Prediction: Random Forest, AdaBoost, and XGBoost

Beny Beny[1,*], Herti Yani[2], Gangga Ramadhan Putra Yupu[3]

[1,3] Department of Informatics, Faculty of Computer Science, Dinamika Bangsa University, Jambi, Indonesia
[2] Computer Science Doctoral, Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Indonesia

ARTICLE INFO

History of the article:
Received August 1, 2025
Revised September 5, 2025
Accepted January 27, 2026

Keywords:
Heart Failure Prediction
Ensemble Learning
Random Forest
AdaBoost
XGBoost

Correspondence:

E-mail: beny@unama.ac.id

ABSTRACT

Heart failure remains a major global health challenge, and early prediction is essential for improving patient outcomes. This study evaluates three ensemble learning methods, namely Random Forest, AdaBoost, and XGBoost, using the Heart Failure Prediction dataset containing 918 patient records from Kaggle. A quantitative experimental design was applied, including preprocessing with KNN imputation, model development, and evaluation using 10-Fold Cross Validation. Performance was assessed through accuracy, precision, recall, F1-score, and AUC-ROC. Random Forest achieved the highest accuracy (0.868), recall (0.907), F1-score (0.884), and AUC-ROC (0.922), while AdaBoost produced the highest precision (0.874). Although the models showed generally similar performance patterns, statistical tests revealed notable distinctions: RF vs. XGB exhibited significant differences in Recall ($p = 0.011$) and F1-score ($p = 0.016$), and the Friedman test identified a significant difference in Recall ($p = 0.034$) across the three models. Feature importance analysis showed that the models consistently emphasized clinically relevant variables, with ST-segment slope, Oldpeak, and exercise-induced angina appearing among the most influential predictors. These features align with recent cardiovascular evidence identifying exercise ECG indicators and stress-response variables as strong predictors of cardiac risk. Overall, the results suggest that recall-related behaviour is the main performance differentiator among the ensemble models, with Random Forest providing a modest advantage in identifying true heart failure cases. The study is limited by its reliance on a single dataset and a relatively small sample size, which may restrict the generalizability of the findings.

1. INTRODUCTION

Heart failure is a major public health concern and remains one of the leading causes of mortality in Indonesia (Sarastri et al., 2023). It occurs when the heart is unable to pump blood effectively, leading to reduced oxygen and nutrient distribution across the body. Several risk factors such as hypertension, diabetes, obesity, smoking, and physical inactivity contribute to its development (Lumi et al., 2021). Early detection plays a critical role in reducing complications and improving survival rates, as timely intervention allows clinicians to deliver appropriate treatment before the condition worsens.

In recent years, machine learning has been widely adopted to support clinical decision-making by transforming complex patient data into predictive insights. While early research focused heavily on conventional single-model algorithms, there is a growing shift toward ensemble learning methods. These methods are particularly valuable because they typically offer improved generalization, reduced variance, and stronger predictive stability compared with single-model approaches. However, despite the potential of these techniques, there remains a lack of comprehensive evaluation comparing modern ensemble frameworks specifically Random Forest, AdaBoost, and XGBoost within a unified clinical context.

This lack of comparative evidence creates a gap in understanding how these methods behave relative to one another, particularly in terms of consistency, robustness, and discriminative ability in real-world clinical data. Therefore, this study aims to address this identified gap by (1) comparing the predictive performance of Random Forest, AdaBoost, and XGBoost using multiple evaluation metrics; (2) assessing the statistical significance of performance differences among these models; and (3) providing empirical evidence on the suitability of these ensemble learning techniques for heart failure risk prediction. By clarifying the strengths and limitations of these models, this study contributes to a more informed selection of machine learning approaches for supporting early heart failure detection, ultimately aiding in more reliable clinical decision support systems.

2. LITERATURE REVIEW

This section reviews the trajectory of machine learning in heart failure prediction, moving from traditional classifiers to modern ensemble techniques.

2.1 Traditional Machine Learning in Clinical Prediction

Early research in heart failure detection relied heavily on conventional, single-model algorithms. Various studies have reported a wide range of accuracy results using Naive Bayes (Barus et al., 2023; Harada et al., 2021; Subarkah et al., 2022) and Support Vector Machines (Arifuddin et al., 2024; Ghasemi & Sharifi, 2025). Other frequently explored methods include K-Nearest Neighbor (Bah, 2022; Kunjachen & Kavitha, 2022; Rahmat et al., 2021; Yunus et al., 2021), Random Forest (Pal & Parija, 2021; Sitanggang & Sitompul, 2024; Tamba, 2022), Logistic Regression Zulkiflee & Rusiman, 2021) and artificial neural network (Al-Jalil & Abu-Naser, 2023; Le et al., 2020). While these models provided a foundational understanding of data-driven diagnosis, their performance often fluctuates depending on the clinical context and the complexity of the feature set, often struggling with high-dimensional or non-linear clinical data.

2.2 Comparative Studies and Performance Benchmarks

A critical trend in the literature is the comparison of different algorithms to identify the most reliable predictor. For instance, Adi and Wintarti (2022) compared SVM, Random Forest, and KNN, reporting that SVM and Random Forest achieved a high accuracy of 97%. Similarly, Hermawan et al. (2024) observed that SVM outperformed Logistic Regression with an accuracy of 85%. More recently, Desiani et al. (2025) evaluated AdaBoost against logistic regression, finding that the boosting approach achieved a superior accuracy of 90%. These comparisons highlight a consistent trend: ensemble-based or non-linear models typically outperform basic statistical classifiers in heart failure tasks.

2.2 Ensemble Methods in Heart Failure Prediction

Ensemble learning has gained prominence due to its ability to combine multiple weak learners into a single strong predictor. Techniques such as Random Forest utilize bagging and feature randomness to reduce overfitting (Rahayu et al., 2020; Rahmah et al., 2023). Conversely, boosting methods like AdaBoost and XGBoost iteratively correct misclassifications, making them highly effective for complex medical datasets (Mahesh et al., 2022; Raj et al., 2024).

Despite the individual success of these models, there is a notable research gap regarding their relative performance when compared within the same experimental framework. Most existing studies focus on comparing an ensemble model against a traditional one, rather than comparing the nuances between different types of ensembles like bagging (Random Forest) and gradient boosting (XGBoost and AdaBoost). This study seeks to bridge that gap by providing a rigorous, multi-metric comparison of these three advanced ensemble techniques.

3. RESEARCH METHODS

This study involved several key stages (Figure 1), including data collection, data preprocessing, model training and testing, and finally, model evaluation.

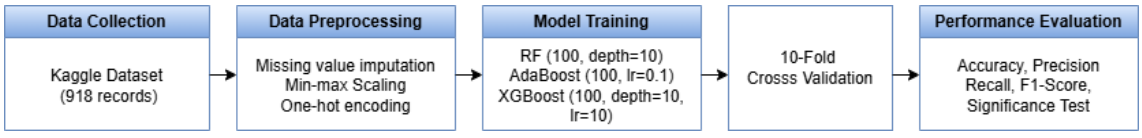

Figure 1. Proposed Methods

3.1. Data Collection Process

The dataset used in this study was obtained from the publicly available Kaggle platform and consisted of 918 patient records with 12 features (Hossain et al., 2024). The dataset includes a mix of data types: numerical attributes (Age, RestingBP, Cholesterol, MaxHR, Oldpeak) stored as integers or floats, and categorical attributes (Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST_Slope) stored as object typesIt includes various medical parameters such as blood pressure, cholesterol levels, and medical history.

3.2. Data Preprocessing

A series of preprocessing steps was conducted to ensure data quality and improve model performance. Although the dataset contained no explicit null values, the Cholesterol feature included a substantial number of zero entries, which are medically implausible and indicate hidden missingness. To address this, K-Nearest Neighbors (KNN) imputation was applied to replace zero Cholesterol values and any other anomalous numerical entries using patterns inferred from similar samples in the dataset. KNN imputation was chosen because it preserves local data structure and produces more realistic estimates than simple mean substitution. Categorical attributes with missing or inconsistent values were imputed using the mode to maintain coherence with their observed distributions. Following imputation, all numerical features were normalized using Min–Max scaling to transform them into a 0–1 range. This was particularly important for boosting-based algorithms such as AdaBoost and XGBoost, which can become disproportionately influenced by features with larger numerical scales. Categorical variables were converted using one-hot encoding to enable the ensemble models to process them effectively without imposing an artificial ordinal structure.

To ensure reproducibility, all preprocessing procedures and the train–test split were performed using a fixed random_state = 42. Additionally, the dataset was evaluated for potential class imbalance; although the class distribution showed moderate imbalance, no resampling techniques (such as SMOTE or undersampling) were applied to avoid introducing synthetic bias. Instead, the imbalance was addressed through cross-validation and the use of evaluation metrics such as F1-score and AUC, which better reflect model performance under uneven class distributions.

3.3. Model Training and Testing

This study evaluates three ensemble learning algorithms, namely Random Forest, AdaBoost, and XGBoost to predict heart failure. The selection of ensemble learning methods over single classifiers is based on their ability to address the inherent complexities of clinical datasets. Traditional single classifiers, such as Decision Trees or Logistic Regression, often struggle with the "bias-variance tradeoff." Single models are frequently prone to overfitting (high variance) when the data is noisy or underfitting (high bias) when the relationships between features are non-linear (Chaithra et al., 2024). Model evaluation was conducted using 10-fold cross-validation, ensuring that each data sample is used for both training and validation (Dutschmann et al., 2023). All models were trained using the same configuration and evaluation procedure to maintain experimental consistency. The experiments were conducted in the Google Colab environment, utilizing Python 3.10 with machine learning libraries including Scikit-learn 1.3, XGBoost 1.7, NumPy, and Pandas. The Colab runtime provided GPU/CPU acceleration and ensured reproducible execution across all training processes.

3.3.1. Random Forest

Random Forest is an ensemble algorithm introduced by Breiman (Wallace et al., 2023). It constructs multiple decision trees using bootstrap sampling and random feature selection, and the final prediction is produced through majority voting. This strategy reduces variance and improves robustness compared to single decision trees. In this study, the Random Forest model was configured with 100 estimators (n_estimators = 100) and a maximum depth of 10 (max_depth = 10), with random_state = 42 to ensure reproducibility. These parameters were selected to balance model complexity and generalization while maintaining stable performance during cross-validation.

3.3.2. AdaBoost

AdaBoost (Adaptive Boosting), introduced by Freund and Schapire (Jasim et al., 2024), builds a sequence of weak learners in which each learner focuses on correcting the errors of its predecessors. The final prediction aggregates all learners using weighted voting based on their accuracy. In this study, the AdaBoost configuration consisted of 100 estimators (n_estimators = 100) and a learning rate of 1.0 (learning_rate = 1.0), with random_state = 42 for reproducibility. This configuration was intended to create a strong boosted ensemble while controlling the amplification of misclassification errors.

3.3.3. XGBoost

XGBoost is an optimized gradient-boosting framework that incorporates regularization and efficient tree construction (Price et al., 2022). It builds decision trees sequentially using gradient-based optimization of an objective function. In this study, the XGBoost model was configured with 100 estimators (n_estimators = 100), a maximum depth of 10 (max_depth = 10), and a learning rate of 0.1 (learning_rate = 0.1), alongside random_state = 42 to ensure consistent results. These settings were chosen to balance model flexibility, computational efficiency, and overfitting control during cross-validation.

3.4. Model Performance Evaluation

A comprehensive evaluation of the model performance was conducted using accuracy, precision, recall, F1-score, and AUC-ROC metrics to assess the overall performance of each algorithm. Accuracy represents the proportion of correct predictions to the total number of predictions and is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

Precision measures the accuracy of the model's positive predictions, indicating the proportion of correct positive predictions. It is formulated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall (or sensitivity) measures the ability of the model to detect all true positives and is formulated as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

The F1-score, which is the harmonic mean of the two, was used to balance precision and recall. It is formulated as follows:

$$F1\text{-}Score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) generally describes the probability that the model will yield a higher classification score for positive samples than for negative samples. AUC values close to 1 indicate excellent classification ability, whereas values close to 0.5 indicate near-random performance.

To determine whether the observed performance differences between the three models were statistically significant, two forms of significance testing were applied. First, pairwise t-tests were conducted on cross-validation scores to evaluate whether the mean performance difference between two algorithms was statistically distinguishable from zero. This test provides insight into whether one classifier consistently outperformed another across folds. Second, the Friedman test, a non-parametric statistical test recommended for machine learning model comparison, was used to assess overall differences among all three classifiers simultaneously. The Friedman p-value indicates whether at least one model performs significantly differently across the set of evaluation metrics. Together, these significance tests strengthen the reliability of the performance comparison by ensuring that the reported differences are not due to random variation in the data.

4. RESULTS AND DISCUSSION

4.1. Exploratory Data Analysis

The exploratory analysis began with a structural and descriptive summary of the dataset, which contains 918 observations and 12 features with no formally missing entries (Table 1). The dataset includes a mix of data types: numerical attributes (Age, RestingBP, Cholesterol, MaxHR, Oldpeak) stored as integers or floats, and categorical attributes (Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST_Slope) stored as object types. This heterogeneity is important because tree-based ensemble models can

natively accommodate mixed data types, but algorithms such as AdaBoost and XGBoost still require categorical values to be transformed into a numerical representation.

Table 1. Dataset Features and Data Types

Column	*Non-Null Count*	*Dtype*
Age	918 non-null	int64
Sex	918 non-null	object
ChestPainType	918 non-null	object
ChotingBP	918 non-null	int64
RestingEBS	918 non-null	object
MaxHR	918 non-null	int64
ExerciseAngina	918 non-null	object
RestingECG	918 non-null	floact
Oldpeak	918 non-null	objet64
ST_Slope	918 non-null	object
HeartDisease	918 non-null	int64

The descriptive statistics as shown in Table 2. further revealed that numerical features exhibited varying scales and distributions: Age (mean ≈ 53.5), RestingBP (mean ≈ 132.4), and MaxHR (mean ≈ 136.8) showed moderate variance, while Cholesterol demonstrated large dispersion (std ≈ 100.2) and an unusually high frequency of zero values suggesting implicit missingness. The target variable (HeartDisease) was moderately balanced (55% positive), enabling fair comparative evaluation across Random Forest, AdaBoost, and XGBoost. These structural and statistical properties guided the preprocessing strategy and shaped expectations regarding each model's sensitivity to skewed, noisy, and mixed-type features.

Table 2. Descriptive Statistics for Dataset Attributes

Metric	Age	Resting BP	Cholesterol	Fasting BS	Max HR	Old peak	Heart Disease
mean	53.511	132.375	198.799	0.233	136.809	0.887	0.553
std	9.432	18.514	109.384	0.423	25.460	1.066	0.497
min	28.000	0.000	0.000	0.000	60.000	-2.600	0.000
25%	47.000	120.000	173.250	0.000	120.000	0.000	0.000
50%	54.000	130.000	223.000	0.000	138.000	0.600	1.000
75%	60.000	140.000	267.000	0.000	156.000	1.500	1.000
max	77.000	200.000	603.000	1.000	202.000	6.200	1.000

As shown in Figure 3, numerical features such as Age and MaxHR were found to follow near-normal distributions. Although tree-based ensemble methods do not require normality, such distributions can still contribute to more stable training dynamics. AdaBoost, in particular, is sensitive to noisy features and irregular feature scales, meaning that near-normal distributions help reduce the risk of error amplification across boosting iterations. In contrast, other numerical attributes most notably RestingBP and Oldpeak exhibited right-skewed distributions. This skewness affects each model differently: Random Forest is largely robust due to its variance-reducing bagging mechanism, although extreme values can still influence individual tree splits. AdaBoost is more susceptible to skew because extreme feature values tend to cause repeated misclassification, thereby increasing their influence as boosting progresses. XGBoost, which relies on gradient-based split optimization, may overfit skewed features if they dominate early split gains. These observations motivated the use of Min–Max scaling to stabilize gradients and control the propagation of biased errors, especially for boosting-based models.

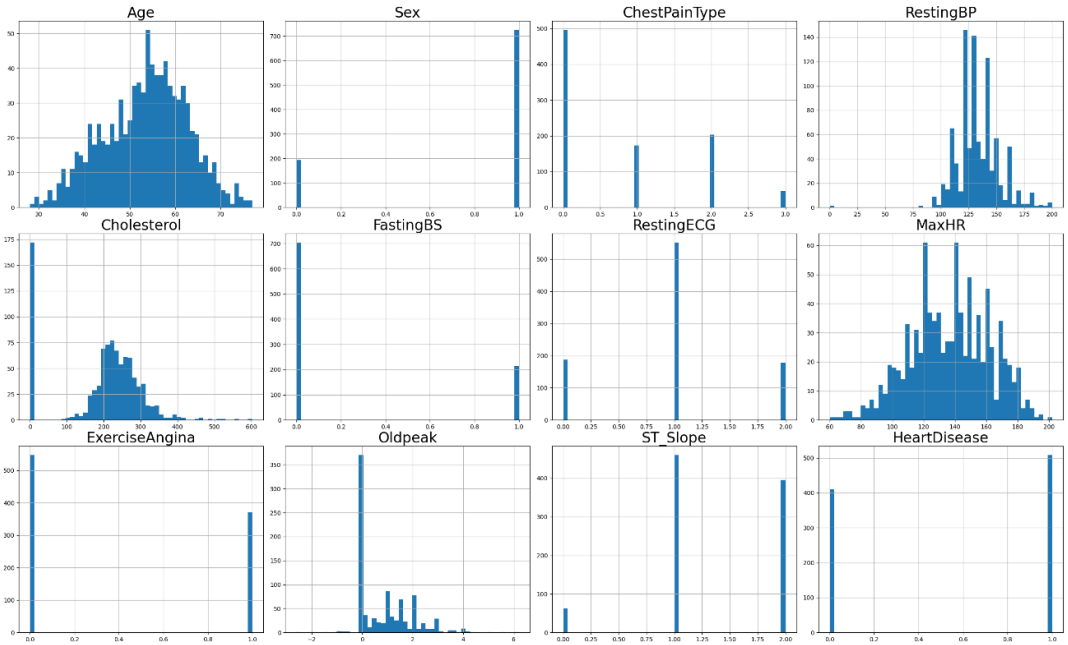

Figure 3. Data distribution of heart disease dataset

A particularly important issue arose in the Cholesterol feature, which displayed an unusually high frequency of zero values. Because zero cholesterol is medically implausible, these values were interpreted as hidden missingness rather than true measurements. This type of noise can perturb model learning in different ways. Random Forest, due to its averaging behavior and random feature sampling, tends to tolerate such irregularities. AdaBoost, however, treats zero values as valid and disproportionately increases the weight of misclassified samples, making it highly vulnerable to this form of hidden noise. XGBoost internally routes missing values through default split directions, yet skewed false-zero patterns may still distort the optimization of its tree structures. To minimize these algorithm-specific risks, the Cholesterol column was processed using KNN imputation, which estimates missing or implausible values based on similarity to neighboring samples. This approach preserves local structure within the data and yields more realistic estimates than simple mean or median substitution, thereby stabilizing the learning dynamics across all three ensemble models.

Boxplot analysis (Figure 4) further revealed substantial variability in features such as Cholesterol. Despite the presence of extreme values, these instances are clinically plausible among cardiac-risk patients and were therefore retained to maintain domain validity. This decision was particularly relevant for XGBoost, whose regularization mechanisms could otherwise prune such values prematurely, potentially reducing model expressiveness.

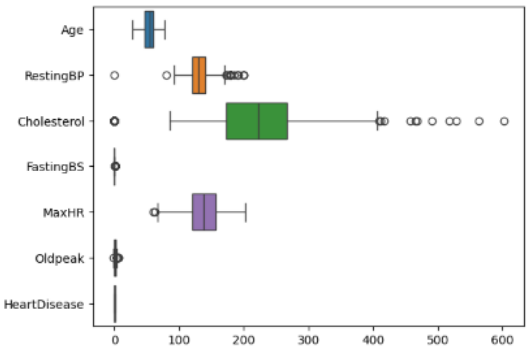

Figure 4. Box plot to visualize outlier

Finally, the correlation matrix (Figure 5) showed strong positive associations between Oldpeak, ExerciseAngina_Y, and ST_Slope_Flat with the target variable, whereas MaxHR exhibited a negative correlation. These relationships align with known medical evidence and suggest high predictive value. From a modeling perspective, Random Forest is expected to leverage these interactions through its capacity to capture nonlinear patterns; AdaBoost benefits from strongly discriminative binary variables such as ExerciseAngina_Y; and XGBoost is positioned to model both linear and complex nonlinear relationships through gradient-optimized splits.

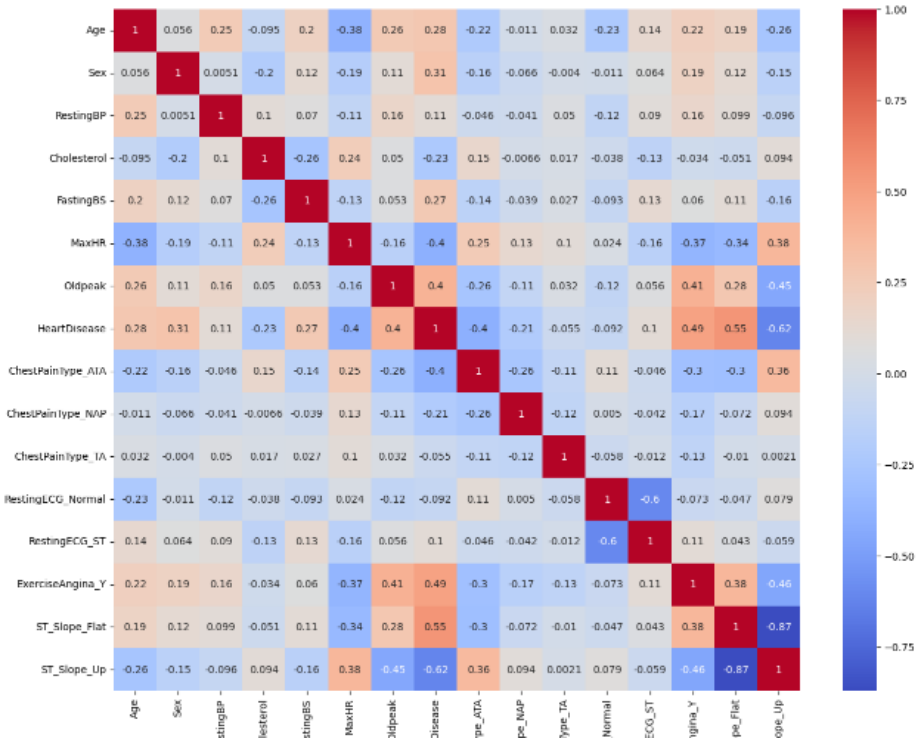

Figure 5. Matrix relationship heatmap

4.2. Feature Importance Analysis

The feature importance analysis reveals distinct weighting patterns across the three ensemble models. In Random Forest, importance values are distributed across several key predictors, with ST_Slope_Up, Oldpeak, ST_Slope_Flat, ExerciseAngina_Y, and MaxHR receiving the highest contributions (Figure 6 (a)).

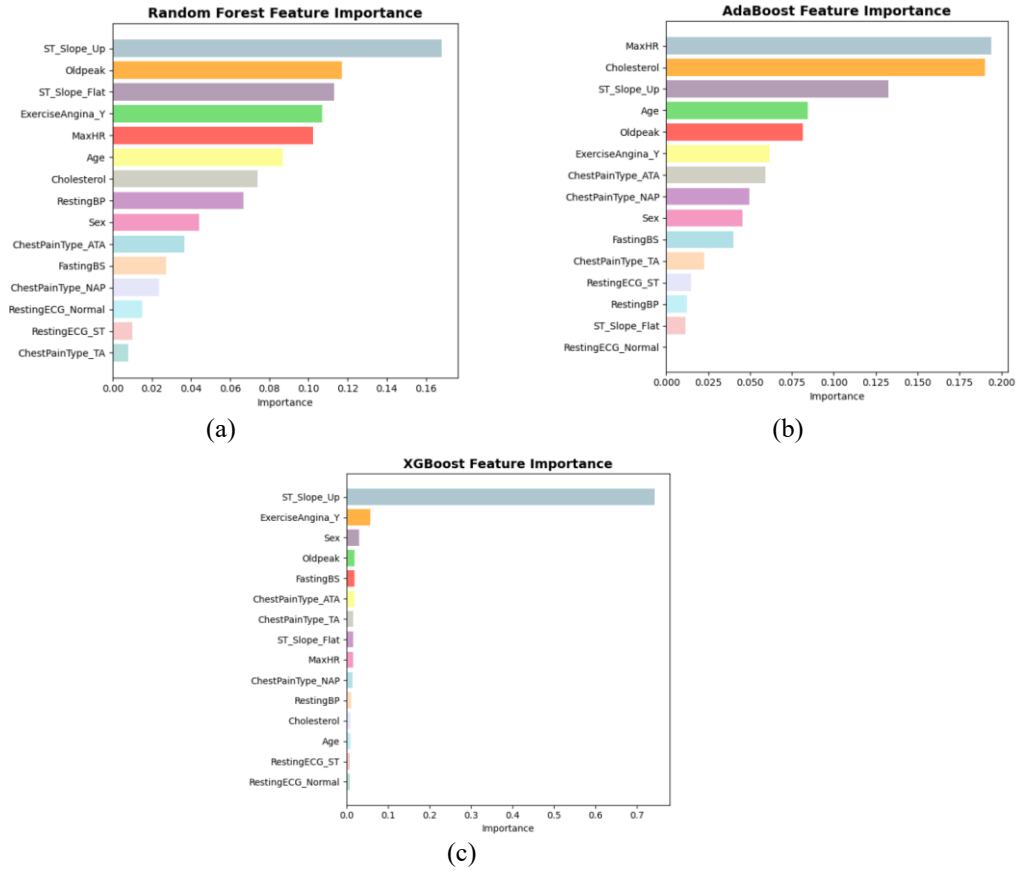

Figure 6. Feature Importance of (a) Random Forest, (b) AdaBoost, and (c) XGBoost

This balanced distribution suggests that Random Forest relies on a broader set of physiological indicators, reflecting its tendency to average information across multiple trees. AdaBoost shows a different pattern: its highest-weighted features include MaxHR, Cholesterol, ST_Slope_Up, Age, and Oldpeak, indicating sensitivity to continuous variables that exhibit moderate separability (Figure 6 (b)). XGBoost displays a markedly sharper concentration of importance on ST_Slope_Up, which dominates the model with a much higher score than other features (Figure 6 (c)). This narrow distribution suggests that XGBoost identifies ST_Slope_Up as the most informative splitting variable within its gradient-boosted structure.

From a clinical perspective, these dominant features align with established cardiovascular literature. ST_Slope and Oldpeak are direct indicators of myocardial ischemia observed during exercise ECG and are strongly associated with obstructive coronary artery disease and elevated risk of cardiac events (Fitzgerald et al., 2022). Exercise-induced angina, MaxHR, and Age are routinely reported as major predictors of cardiovascular stress response and functional capacity. Elevated Cholesterol is also a well-known metabolic risk factor in heart failure progression (Stretti et al., 2021). The models' reliance on these features is therefore consistent with prior medical findings and previous machine-learning studies on heart disease prediction, which frequently highlight ST-segment behavior and stress-test indicators as among the strongest predictors.

4.3. Model Performance

The evaluation of the three ensemble learning methods using 10-fold cross-validation showed that all models produced competitive results across the primary metrics (Table 3). Random Forest achieved the highest accuracy (0.868), recall (0.907), and AUC-ROC (0.922), indicating that it identified a larger proportion of true heart-disease cases and exhibited stronger class separability for this dataset. AdaBoost, in contrast, obtained the highest precision (0.874), reflecting its tendency to produce fewer false-positive predictions. XGBoost recorded slightly lower accuracy (0.848) and precision (0.850) than the other two models, but its recall (0.884) and AUC-ROC (0.917) remained within a close range. Across all three methods, the F1-scores 0.884, 0.877, and 0.866 were relatively similar, suggesting comparable balance between sensitivity and specificity.

Table 3. Model performance comparison on 10-Fold Cross Validation

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	0.868	0.864	0.907	0.884	0.922
AdaBoost	0.863	0.874	0.882	0.877	0.918
XGBoost	0.848	0.850	0.884	0.866	0.917

A closer examination of the metric patterns highlights differences in how each model responds to the characteristics of the dataset. Random Forest's higher recall suggests that its bagging-based structure tends to identify a larger portion of positive cases under the given feature distributions. AdaBoost's higher precision is consistent with its boosting mechanism, which adjusts sample weights across iterations and can lead to fewer false-positive classifications. XGBoost shows slightly lower values on some metrics, reflecting its gradient-based optimization process, which models feature interactions differently from the other two methods. Overall, the results indicate that the three ensemble algorithms prioritize different aspects of the classification task, producing performance profiles that differ in emphasis but remain numerically close.

To determine whether these performance differences were statistically meaningful, pairwise t-tests and a Friedman test were conducted (Table 4). Most pairwise t-test p-values exceeded the 0.05 significance threshold, indicating that the differences in Accuracy, Precision, and AUC among the models were not statistically significant. Two comparisons, however, did reach significance: RF vs. XGB in Recall ($p = 0.011$) and RF vs. XGB in F1-score ($p = 0.016$), suggesting measurable differences for these specific metrics. The Friedman test also showed no significant differences for Accuracy, Precision, F1-score, or AUC, but it identified a significant difference in Recall ($p = 0.034$) across the three models. Overall, the statistical tests indicate that while some metric-level differences exist particularly in recall-related behavior the three ensemble methods exhibit broadly similar performance patterns within the dataset.

Table 4. Performance Significant Test

Metric	RF vs AB t-test	RF vs XGB t-test	AB vs XGB t-test	Friedman P-Value
Accuracy	0.697	0.023	0.220	0.067
Precision	0.385	0.108	0.043	0.103
Recall	0.027	0.011	0.772	0.034
F1	0.484	0.016	0.348	0.128
AUC	0.468	0.417	0.983	0.583

5. CONCLUSIONS AND RECOMMENDATIONS

This study compared the performance of three ensemble learning algorithms, namely Random Forest, AdaBoost, and XGBoost for heart failure prediction using 10-fold cross-validation. Random Forest achieved the highest accuracy (0.868), recall (0.907), F1-score (0.884), and AUC-ROC (0.922), while AdaBoost obtained the highest precision (0.874). XGBoost produced values that were close to those of the other two models across all evaluation metrics. Overall, the statistical analyses confirm that the three ensemble models achieve broadly comparable performance, with no significant differences for Accuracy, Precision, or AUC. However, two pairwise comparisons RF vs. XGB in Recall and F1-score showed statistically significant differences, and the Friedman test also indicated a significant difference in Recall across the models. These findings highlight that while overall performance levels are similar, recall-related behavior distinguishes the models most clearly.

This study is limited by the use of a single dataset with a moderate sample size, which may constrain the generalizability of the findings. Additionally, the models were evaluated using baseline hyperparameters, and further tuning could influence comparative outcomes. Future research may include testing additional algorithms such as deep learning methods, applying broader hyperparameter optimization, incorporating interpretability techniques such as SHAP, and validating the models on larger or external datasets to better understand model behavior and robustness.

REFERENCES

Adi, S., & Wintarti, A. (2022). Komparasi metode support vector machine (SVM), K-Nearest Neighbors (KNN), Dan Random Forest (RF) untuk prediksi penyakit gagal jantung. *MATHunesa: Jurnal Ilmiah Matematika*, *10*(2), 258–268.

Al-Jalil, K. M. A., & Abu-Naser, S. S. (2023). *Artificial Neural Network Heart Failure Prediction Using JNN*.

Arifuddin, A., Buana, G. S., Vinarti, R. A., & Djunaidy, A. (2024). Performance comparison of decision tree and support vector machine algorithms for heart failure prediction. *Procedia Computer Science*, *234*, 628–636.

Bah, I. (2022). Knn algorithm used for heart attack detection. *FES Journal of Engineering Sciences*, *11*(1), 7–19.

Barus, O. P., Lauwren, K., & Pangaribuan, J. J. (2023). Implementation of the Naive Bayes Algorithm to Predict the Safety of Heart Failure Patients. *IAIC International Conference Series*, *4*(1), 172–177.

Chaithra, K. N., Shukla, A., Kanodiya, H., & Ray, S. (2024). Performance of XG Boost over other ML models for Prediction of CVD. *2024 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, 1–6.

Desiani, A., Amran, A., Andriani, Y., Wahyuni, T., & Rizki, F. (2025). Perbandingan Algoritma Logistic Regression Dan Adaptive Boosting (Adaboost) Dalam Klasifikasi Penyakit Gagal Jantung. *Jurnal Teknologi Informasi: Jurnal Keilmuan Dan Aplikasi Bidang Teknik Informatika*, *19*(1), 72–78.

Dutschmann, T.-M., Kinzel, L., Ter Laak, A., & Baumann, K. (2023). Large-scale evaluation of k-fold cross-validation ensembles for uncertainty estimation. *Journal of Cheminformatics*, *15*, 49.

Fitzgerald, B. T., Smith, E., & Scalia, G. M. (2022). What are the prognostic implications and factors relating to exercise induced electrocardiographic ST segment changes in the setting of a non-ischemic stress echocardiogram? *International Journal of Cardiology*, *364*, 157–161.

Ghasemi, F., & Sharifi, S. (2025). Heart Failure Prediction Using Support Vector Machine. *International Journal of Novel Research in Life Sciences*.

Grgić, V., Mušić, D., & Babović, E. (2021). Model for predicting heart failure using Random Forest and Logistic Regression algorithms. *IOP Conference Series: Materials Science and Engineering*, *1208*(1), 012039.

Harada, D., Asanoi, H., Noto, T., & Takagawa, J. (2021). Naive Bayes Prediction of the Development of Cardiac Events in Heart Failure With Preserved Ejection Fraction in an Outpatient Clinic―Beyond B-Type Natriuretic Peptide―. *Circulation Journal*, *86*(1), 37–46.

Hermawan, K. A., Rizki, A., Sinaga, D. K., & Suwarman, H. R. (2024). Prediksi Gagal Jantung Berbasis Machine Learning Menggunakan Support Vector Machine dan Regresi Logistik. *Seminar Nasional Penelitian (SEMNAS CORISINDO 2024)*, 436–441.

Hossain, M. A., Shawkat, S. Bin, Sharif, K. S., Hossain, M. I., Asmani, H., & Rahman, M. M. (2024). Precisioncardio: A comprehensive machine learning approach for accurate prediction of heart failure trajectory. *2024 IEEE 30th International Conference on Telecommunications (ICT)*, 1–4.

Jasim, A. A., Hazim, L. R., Mohammedqasim, H., Mohammedqasem, R., Ata, O., & Salman, O. H. (2024). e-Diagnostic system for diabetes disease prediction on an IoMT environment-based hyper AdaBoost machine learning model. *The Journal of Supercomputing*, *80*(11), 15664–15689.

Kunjachen, L. M., & Kavitha, R. (2022). Comparative Study on Detection of Heart Disease using KNN Algorithm. *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, 14–19.

Le, M. T., Vo, M. T., Mai, L., & Dao, S. V. T. (2020). Predicting heart failure using deep neural network. *2020 International Conference on Advanced Technologies for Communications (ATC)*, 221–225.

Lumi, A. P., Joseph, V. F. F., & Polii, N. C. I. (2021). Rehabilitasi jantung pada pasien gagal jantung kronik. *Jurnal Biomedik: JBM*, *13*(3), 309–316.

Mahesh, T. R., Dhilip Kumar, V., Vinoth Kumar, V., Asghar, J., Geman, O., Arulkumaran, G., & Arun, N. (2022). AdaBoost ensemble methods using K-fold cross validation for survivability with the early detection of heart disease. *Computational Intelligence and Neuroscience*, *2022*(1), 9005278.

Pal, M., & Parija, S. (2021). Prediction of heart diseases using random forest. *Journal of Physics: Conference Series*, *1817*(1), 012009.

Price, J., Yamazaki, T., Fujihara, K., & Sone, H. (2022). XGBoost: interpretable machine learning approach in medicine. *2022 5th World Symposium on Communication Engineering (WSCE)*, 109–113.

Rahayu, S., Purnama, J. J., Pohan, A. B., Nugraha, F. S., Nurdiani, S., & Hadianti, S. (2020). Prediction of survival of heart failure patients using random forest. *Jurnal Pilar Nusa Mandiri*, *16*(2), 255–260.

Rahmah, A., Sepriyanti, N., Zikri, M. H., Ambarani, I., & bin Shahar, M. Y. (2023). Implementation of Support Vector Machine and Random Forest for Heart Failure Disease Classification. *Public Research Journal of Engineering, Data Technology and Computer Science*, *1*(1), 34–40.

Rahmat, D., Putra, A. A., & Setiawan, A. W. (2021). Heart disease prediction using K-nearest neighbor. *2021 International Conference on Electrical Engineering and Informatics (ICEEI)*, 1–6.

Raj, S., Vani, R., Raja, B., Harsha, T., Drakshayani, T., & Charith, R. (2024). HEART DISEASE DETECTION USING XGB-CLASSIFIER AND FAILURE PREDICTION USING GRADIENT BOOSTING. *Journal Of Nonlinear Analysis and Optimization*, *15*.

Sarastri, Y., Zebua, J. I., Lubis, P. N., Zahra, F., & Lubis, A. C. (2023). Admission hyponatraemia as heart failure events predictor in patients with acute heart failure. *ESC Heart Failure*, *10*(5), 2966–2972.

Sitanggang, B. F., & Sitompul, P. (2024). Deteksi Awal Kelangsungan Hidup Pasien Gagal Jantung Menggunakan Machine Learning Metode Random Forest. *Innovative: Journal Of Social Science Research*, *4*(2), 3347–3357.

Stretti, L., Zippo, D., Coats, A. J. S., Anker, M. S., von Haehling, S., Metra, M., & Tomasoni, D. (2021). A year in heart failure: an update of recent findings. *ESC Heart Failure*, *8*(6), 4370–4393.

Subarkah, P., Risma, W., & Aditya, R. (2022). Comparison of correlated algorithm accuracy Naive Bayes Classifier and Naive Bayes Classifier for heart failure classification. *Vol*, *14*, 120–125.

Tamba, S. P. (2022). Prediksi Penyakit Gagal Jantung Dengan Menggunakan Random forest. *Jurnal Sistem Informasi Dan Ilmu Komputer*, *5*(2), 176–181.

Wallace, M. L., Mentch, L., Wheeler, B. J., Tapia, A. L., Richards, M., Zhou, S., Yi, L., Redline, S., & Buysse, D. J. (2023). Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. *BMC Medical Research Methodology*, *23*(1), 144.

Yunus, R., Ulfa, U., & Safitri, M. D. (2021). Application of the K-Nearest Neighbors (K-NN) algorithm for classification of heart failure. *Journal of Applied Intelligent System*, *6*(1), 1–9.

Zulkiflee, N. F., & Rusiman, M. S. (2021). Heart Disease Prediction Using Logistic Regression. *Enhanced Knowledge in Sciences and Technology*, *1*(2), 177–184.